# Evaluating User Behavior on Data Collections in a Digital Library

Michalis Sfakakis[1] and Sarantos Kapidakis[2]

[1]National Documentation Centre / National Hellenic Research Foundation
48 Vas. Constantinou, GR-11635 Athens, Greece
msfaka@ekt.gr
[2]Archive and Library Sciences Department / Ionian University
Plateia Eleftherias, Paleo Anaktoro, Corfu 49100, Greece
sarantos@ionio.gr

## Abstract

We evaluate the usage of a Digital Library with many different collections, by examining its log files, and we concluded that the access points that the users mostly refer to, depend heavily on the type of content of the collection. We also found that most users not only tend to use simple query structures (e.g. one search term) and very few operations per session but they also reduce the complexity of their sessions, as they get more experienced.

## Introduction

The evolution of Digital Libraries attends great interest by researchers in a variety of disciplines during the last years. Especially the study to understand and evaluate their usage has become a centric point in a number of Digital Library projects ([8], [9]) and specifies a number of critical factors during the design, creation and development process of a Digital Library ([10]).

Depending on the study and its use, a number of appropriate qualitative or quantitative methods exist ([3], [4]) to accomplish it. An unobtrusive way to study and evaluate user behavior is the Transaction Log Analysis. Although log analysis is used as an effective method to assess how users actually interact with a working Digital Library, this method hardly provides any information about the users' reasons behind their specific behavior - which is also very difficult to extract – and it is lack of giving information on their intentions. The accuracy of this quantitative method heavily depends on the detail of the information logged (automatically by the system), the period of time used to log the information, the usage and the number of the performed transactions during the log period. Such data are not usually publicly available (especially in detail) because of privacy constraints. For these reasons and due to that large Digital Libraries have recently started developing, only a few studies exist based on this technique ([5], [6], [7]).

In this work, based on the logged information, we study and evaluate the behavioral tendencies of different user groups on a variety of collections in the Digital Library of the Hellenic National Documentation Centre (*NDC*). The Digital Library of NDC (http://theses.ndc.gr) is one of the most significant in Greece and consists of more than ten collections of diverse types. Most of these collections are unique world wide with internationally interesting content. In particular, the "Hellenic Ph.D. Dissertations Thesis" collection is part of the international Networked Digital Library of Theses and Dissertation Initiative ([2]). The Digital Library of NDC is targeted to a number of diverse types of user groups (e.g. students, researchers, professionals, librarians, etc.), mainly in Greece, from a variety of scientific domains.

In the following section we describe the goal and the methodology of this study. We also describe the collections, their characteristics, the target user groups they refer to and the functionality of the available operations by the system. Then we present some our most important observations from the search operation usage and formulation, the Access Points usage and how users accomplish their requests, together with our interpretation and conclusions. Finally we present a number of interesting issues arrived from this work for further evaluation and research.

## Purpose and Methodology of the study

The goal of this study is to compare and evaluate the differences on the usage among data collections, based on the collection content type, metadata and characteristics and also to approach the way diverse kinds of users accomplish their requests.

For a period of twenty months, we logged the operations performed by the users on the content of many different collections of a Digital Library, using a specific web based retrieval system. Considering the content type (e.g. PhD theses, articles in a specific scientific area, Books and Periodicals Union catalogues etc.), the structure and the quality of the collection, plus the target group they refer to, we selected the ten most used ones and classified them into four categories.

Category one consists of the collections: *Hellenic Ph.D. Dissertations Thesis* (*C1*) and *Hellenic Scientific Libraries Serials Union Catalogue* (*C2*), targeted to diverse kinds of scientific user groups (e.g. students, researchers etc.) from all scientific domains.

Category two consists of the collections: *Medical Bibliography – Hippocrates* (*C3*) and *Social Science Bibliography – GLAFKA* (*C4*), with simple metadata structure, targeted to a specific scientific user group (e.g. doctors, sociologists, researchers).

Category three consists of the collections: *Hellenic Archaeological Records – grARGOS* (*C5*) and *International Archaeological Records – intARGOS* (*C6*), including library material with diverse types of data, targeted to a specific scientific user group (e.g. researchers on Archaeology).

Category four consists of the collections: *Hellenic School Libraries* (*C7*) and *Hellenic Public Libraries Union Catalogue* (*C8*), union catalogs for library materials from many domains, targeted to librarians.

The remaining two collections are the *ARGOS – Serials Union Catalogue* and the *Evonimos Ecological Library*, which enjoy smaller use and we do not examine them separately, for simplicity, but we count their usage on the aggregated results.

All the above collections are structured using the UNIMARC format but they do not use the same detail on metadata description. From their 300,000 metadata records there are links giving online access to 14,000 digitized documents composed of 2,000,000 scanned pages and few other object formats.

The web-based retrieval system that we monitored is implementing a Z39.50 client and connects to a Z39.50 server. The users start their sessions by selecting and connecting to a collection. After connecting to a collection, a user may express his search request or browse specific *Access Points* (e.g. extracts information about metadata indexing - the terminology used for naming them is the one of the Z39.50 attribute set bib-1 as defined in [1]) and then to retrieve (*present*) the documents. In some cases, there is the ability to further access the object (document) that includes the full text, mostly in scanned images. There are seventeen available Access Points and the "search" operation supports Boolean combinations of them. When the user browse the terms from a specific Access Point, the system permits either to select a term in order to use it in a "search" operation or to retrieve ("present") the corresponding documents for display or further processing (searching and retrieving). From the more advanced searching techniques, the system also supports Boolean search combination of previously issued result sets, Search History and Selection of specific records from individual result sets.

The study covering period from August 2000 till March 2002, gave a set of 490,042 operations to process and evaluate (Table 1).

**Table 1.** Summary of Processed Data

| Covering Period | Number of Collections | Number of Sessions | Number of Operations |
|---|---|---|---|
| August 2000-March 2002 (20 Months) | 10 | 64,597 | 490,042 |

Fig. 1 and Fig. 2 depict the number of operations and sessions respectively for all collections on a monthly basis. During the study covering period, no major modifications occurred on the two basic components of the Digital Library, the collections and the retrieval system. On the other hand, the number of operations (Fig. 1), sessions

(Fig. 2) and users increased while maintaining a yearly periodic variation, and the number of sessions and operations seem to have the same transitions.

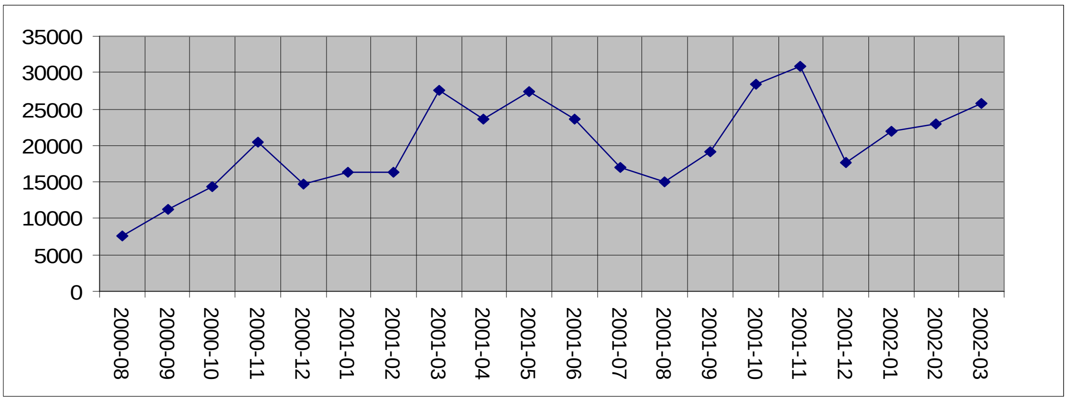

**Fig. 1.** Number of Operations per month

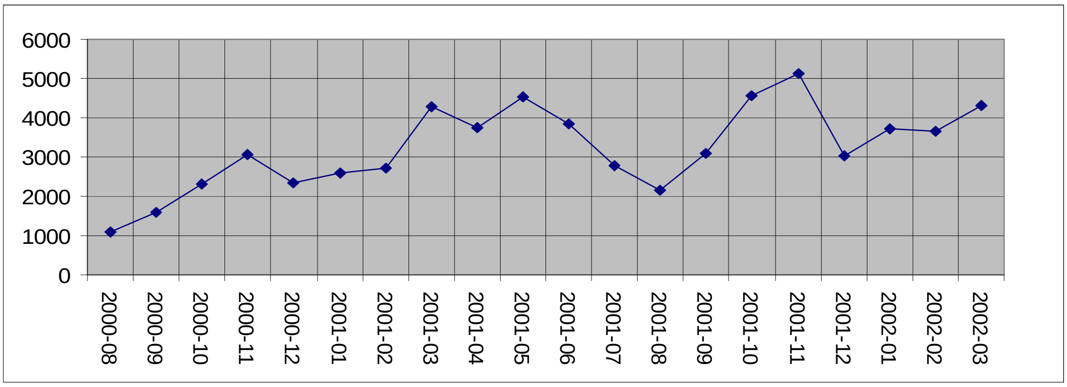

**Fig. 2.** Number of Sessions per month

## Search Operation Usage and Formulation

From the processed logs, as presented on Table 2, we found that the percentage of the search operations is 38.34% of the total number of operations. The majority (81.75%) of these search operations were formulated using one search term (simple query). Finally, the users did not make use of advanced querying techniques (0.1%) by formulating search operations using previously issued result sets.

**Table 2.** Summary of Search use and formulation

| | |
|---|---|
| Total Search operations | 187,898 (38.34% total operations) |
| Search operations with only one search term (Simple Queries) | 153,283 (81.57% Search Operations) |
| Search operations using Boolean expressions (Compound Queries) | 34,615 (18.43% Search Operations) |
| Use of previously Issued result sets in compound queries | 200 (0.1% Search Operations) |

## Use of Access Points Evaluation

Table 3 shows the number of times each Access Point has been used, for each collection and all collections together, and also the Access Points order of preference. The number after the Access Point name is from the Z39.50 bib-1 attribute set.

The evaluation of the usage of Access Points, shows that the most commonly used Access Points, for all the collections in the Digital Library, are the “Any”, “Author”, “Title”, “Subject Heading” (Table 3), from the seventeen ones used in the metadata (the used terminology for naming Access Points is the one used by the Z39.50 attribute set bib-1). The vast majority of all users, independent of user group, used the “Any” Access Point for almost all collections. The only exception occurs at collection C6, were the most used Access Point is the “Author” which could be explained from the specialized subject area of the collection’s content (Archaeological Records) in combination with the specific type of its closed targeted user group’s requests.

Another interesting observation with regard to the first two categories of the collections is the big usage difference (60.5% - 80.9%) between the two most used Access Points. These collections consist of content with simple metadata structure and are targeted to a number of diverse types of occasional users.
At the third category which consists of collections with typical library material (e.g. more complex metadata structure, diverse kinds of material) that impose a more accurate process by professionals with consequence a better quality of metadata and targeted to a more specific user group, there is a balance between the three most used Access Points.

The most balanced usage between the commonly used Access Points, happens at the forth category, which consists of collections with common characteristics as those in the third category and targeted to librarians.

Consequently, we observe that the usage of these commonly used Access Points depends mainly on the collection they belong to as well as on the user group type they are targeted to.

**Table 3.** Summary of Access Points use per Collection

| | Total | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 |
|---|---|---|---|---|---|---|---|---|---|
| Any(1016) | 114,412 | 55,137 | 4,946 | 20,474 | 9,119 | 5,688 | 3,729 | 5,488 | 5,539 |
| Author(1003) | 43,031 | 21,823 | 770 | 1,446 | 1,061 | 4,093 | 4,445 | 4,236 | 4,146 |
| Title(4) | 37,048 | 14,227 | 1,942 | 2,679 | 2,338 | 3,802 | 3,252 | 3,861 | 2,810 |
| Subject Heading(21) | 22,760 | 10,997 | 594 | 1,073 | 657 | 572 | 641 | 3,883 | 2,454 |
| | | Any<br>Author<br>Title<br>Subject | Any<br>**Title**<br>**Author**<br>Subject | Any<br>**Title**<br>**Author**<br>Subject | Any<br>**Title**<br>**Author**<br>Subject | Any<br>Author<br>Title<br>Subject | **Author**<br>**Any**<br>Title<br>Subject | Any<br>Author<br>**Subject**<br>**Title** | Any<br>Author<br>Title<br>Subject |

Table 4 displays the usage of Access Point combinations for each collection and all collections together. We first observe that the Access Point “Any” is not that dominant in Access Point combinations as it was in single Access Point specifications. We also observe that the difference between the two most used Access Point combinations follows the previously observed Access Point usage pattern. Finally, for the majority of the collections, the most commonly used combination of Access Points is the “Title-Any”, except for the collections C5, C6, C8. We have already seen (Table 3) that these collections have a more uniform usage on their single Access Points, without overusing the “Any” Access Point, and consequently the most commonly used combination of Access Points for them is the “Title-Author”.

We also observe that the Access Point “Title” is used much more often on Access Point combinations, although it is the third one in the list of the most used Access Points, which indicate that “Title” is used in more sophisticated “search” operations and by more sophisticated users.

**Table 4.** Summary of Access Points Combination use per Collection

| | Total | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 |
|---|---|---|---|---|---|---|---|---|---|
| Title(4) - Any(1016) | 4,262 | **1,925** | **137** | **640** | **489** | 177 | 236 | **370** | 179 |
| Title(4) - Author(1003) | 2,860 | 776 | 64 | 107 | 43 | **656** | **541** | 294 | **301** |
| Author(1003) - Any(1016) | 1,388 | 823 | 51 | 104 | 41 | 69 | 119 | 54 | 109 |
| Subject Heading(21) – Any(1016) | 976 | 540 | 40 | 86 | 47 | 33 | 37 | 79 | 77 |
| Title(4) – Subject Heading(21) | 602 | 286 | 24 | 33 | 21 | 25 | 35 | 67 | 70 |
| Title(4) - Author(1003) – Any(1016) | 506 | 293 | 18 | 39 | 24 | 22 | 30 | 15 | 50 |
| Subject Heading(21) - Author(1003) | 503 | 264 | 10 | 21 | 10 | 23 | 28 | 60 | 66 |

Comparing the results that, the vast majority (81.57%) of the search queries consist of one search term (Table 2) and most users for almost all collections use a general Access Point ("Any") to accomplish their requests with big usage difference from the next, more specific, Access Point ("Author or "Title""), we can derive that new users will need more operations to accomplish their requests which impacts the increase of the number of operations per session when new users enter the system.

## User Behavior (how user accomplishes the job)

Fig. 3 shows the monthly average operations per session aggregated for all ten collections studied, on a monthly basis. Similar lines correspond to each one of the studied collections.

The average number of operations per session in general drops during the study period. Does this mean that the vast majority of old users becomes more experts and expresses their requests using fewer operations?

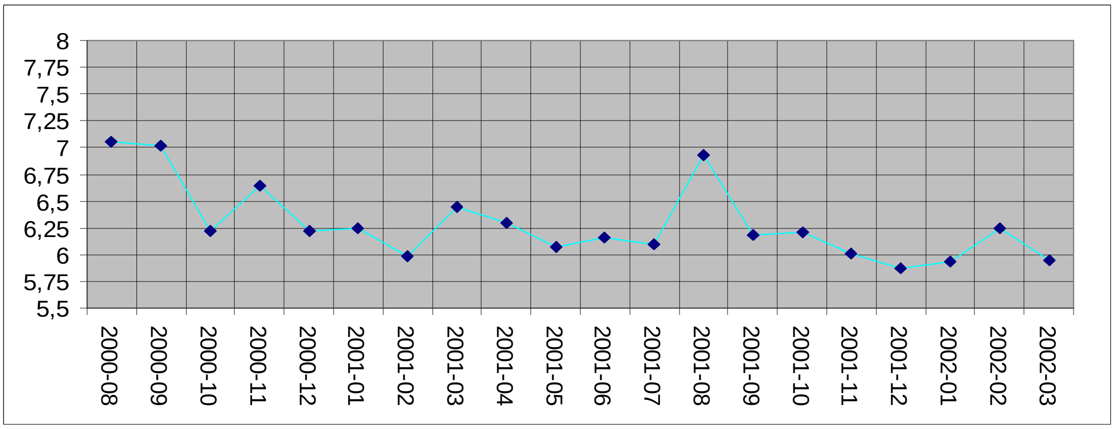

**Fig. 3.** Average Operations per session

Fig. 4 shows the monthly proportion of sessions with operations less than or equal to three per session aggregated for all ten collections studied, on a monthly basis. Similar lines correspond to each one of the studied collections.

We observe that in each month (Fig. 4), three operations are enough to fulfill at least half of the sessions. Also, the number of sessions with less than or equal three operations per session have a constant fluctuation, which indicates that there is a balance on the number of sessions with the same number of operations per session, between new users and old users that become more experienced.

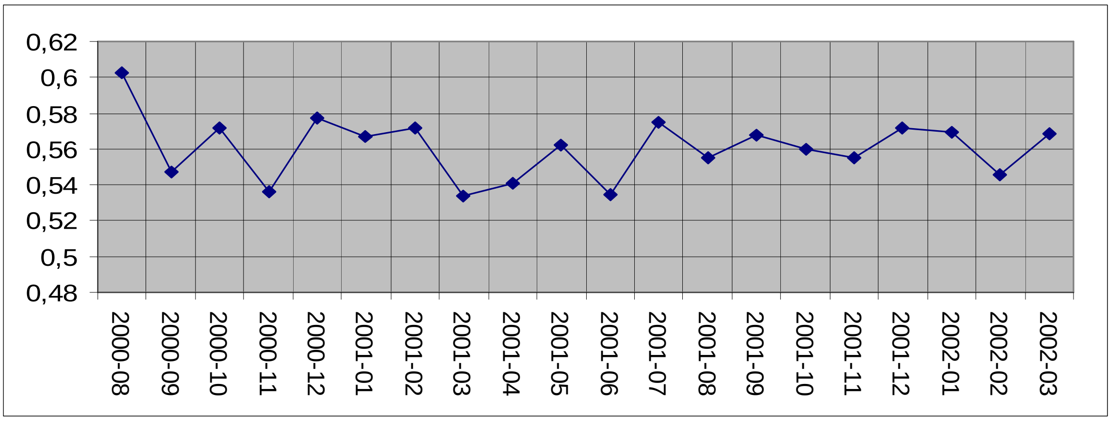

**Fig. 4.** Proportion of Sessions with operations less than or equal 3 per session

Another interesting question is how we measure the experience of the users. The experience of the users will certainly increase by time, but how can we distinguish it from that of newer users, on a system that does not record the identification of the users?
We assume that one aspect of the experience of the user is measured by the number of operations that are included in a session, the full sequence of operations that the user performed. We have already concluded that most users perform few operations in order to find their material, but as the users become more experienced, do they use more operations (been able to make more complex sessions) or less operations (been more specific and efficient) in their sessions? The addition of new users into the system makes the distinction more difficult.

Fig. 5 shows the number of sessions for each number of operations (from 1 to 30) per session for five representative months, aggregated for all collections. From fig. 5 we can see that on the later stages in our Digital Library lifetime, the increased number of users corresponds to only an increase to the number of sessions that have only one operation. We already observed, on the evaluation of Access Points usage, that new users perform queries with many operations per session.  We also believe that it is unlikely that all new users perform only queries with one operation per session, while we can see from fig. 2 that the total number of sessions in the last three of the depicted months are practically the same, so we conclude that older users decrease the number of operations into their sessions, in a way that (by coincidence) corresponds to or outperforms the increase of new users performing the same number of operations per session.

Thus expert users use fewer operations per session than non-expert users, and the users decrease the number of operations in their sessions during the time they use the Digital Library.

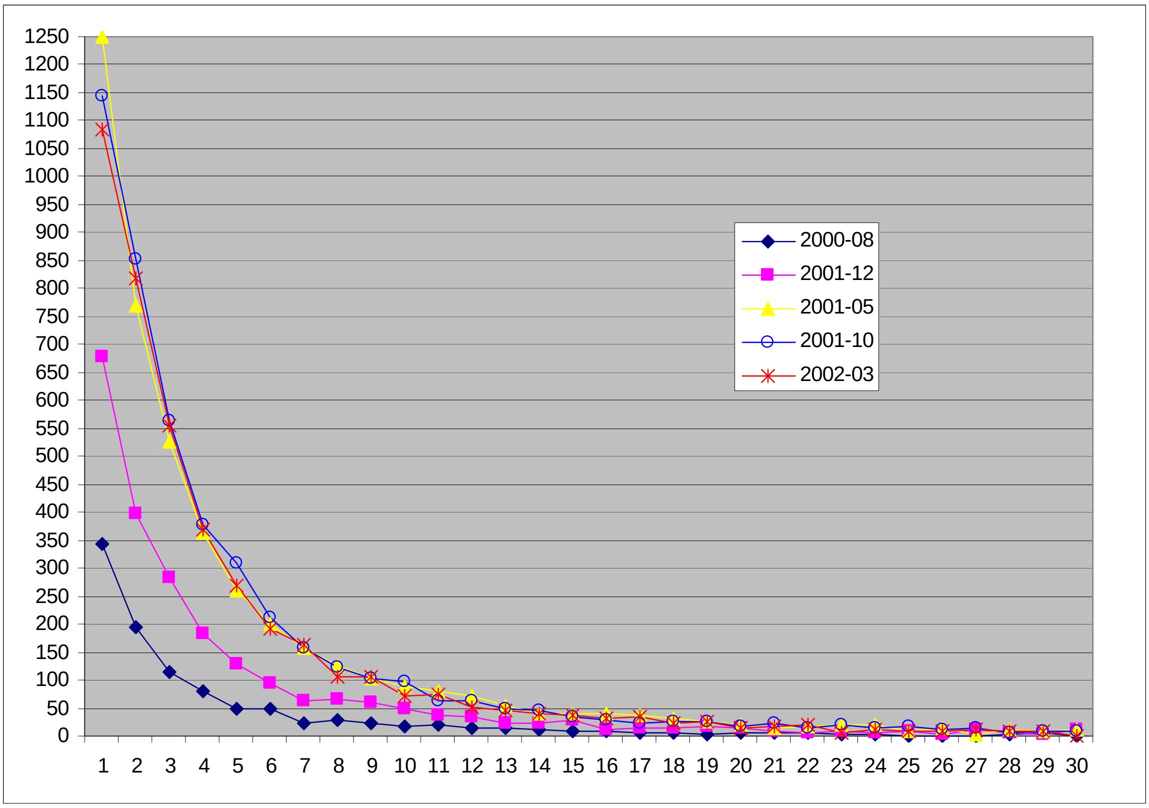

**Fig. 5.** Number of operations per session - Number of Sessions

## Conclusions and Future Research

We studied the Access Points usage and we derived that the “Any” Access Point is used by novice or non specialized users, while other Access Points, like “Title” and “Author”, are mostly in use by experienced and sophisticated users, on complex queries and on collections with more complex metadata. We also examined the number of operations and sessions and we concluded that expert users tend to decrease the number of operations in their sessions, by been more explicit and efficient.

From this work a number of interesting points arrives for future evaluation and research. How does the Access Points usage evolve by the time? A more detailed analysis for the search term formulation (e.g. word, phrase, truncation) used by the same group of users to accomplish their search requests per collection would be interesting. How previously issued user behavior results, differentiated per collection? Another point of interest is how different user types (e.g. professionals, ordinary users) behave under the same circumstances. What sequences (patterns) of operations (i.e. number of "Presents" follows the "Search" operation, etc.) in sessions do different types of users adopt? Finally the Query formulation complexity progress during the time period is also another interesting point of evaluation.

## References

1. ANSI/NISO: Z39.50 Information Retrieval: application service definition and protocol specification: approved May 10, 1995.
2. E. Fox, Robert Hall, Neill A. Kipp, John L. Eaton, Gail McMillan, and Paul Mather. NDLTD: Encouraging International Collaboration in the Academy. In Special Issue on Digital Libraries, DESIDOC Bulletin of Information Technology (DBIT), 17(6): 45-56, Nov. 1997.
3. Bains S., "End-User Searching Behavior: Considering Methodologies", The Katharine Sharp Review, No. 4, Winter 1997.
4. Covey, D. T., "Usage and Usability Assessment: Library Practices and Concerns", Washington, D.C., Digital Library Federation Council on Library and Information Resources, January 2002, ISBN 1-887334-89-0.
5. Jones, S., Cunningham, S. J., McNab, R. J. and Boddie, S., "A Transaction log Analysis of a digital library", International Journal on Digital Libraries, v. 3:no. 2 (2000), pp. 152-169.
6. Mahoui, M., Cunningham, S. J., "Search Behavior in a Research-Oriented Digital Library", ECDL 2001, LNCS 2163, pp. 13-24.
7. Mahoui, M., Cunningham, S. J., "A Comparative Log Analysis of Two Computing Collections", Research and Advanced Technology for Digital Libraries: Proceedings of the 4th European Conference, ECDL Lisbon, Portugal, Sept. 2000, pp. 418-423.
8. Peterson Bishop, A., "Working toward an understanding of digital library use: a report on the user research efforts of the NSF/ARPA/NASA DLI projects", D-Lib Magazine, October 1995.
9. Payette, S.D. and Rieger, O.Y. "Z39.50 The User's Perspective", D-Lib Magazine, April 1997.
10. Van House, N.A. et. al., "User centered iterative design for digital libraries: the Cypress experience", D-Lib Magazine, February 1996.